Methods for Classifying Changes in Bacterial Prevalence Over Time

Raymond G Hoffmann[1], Ke Yan[2], Sandra McLellan[3], Jessica VandeWalle[4]

[1]Pediatrics and The Children's Research Institute, Medical College of Wisconsin
[2]Pediatrics and The Children's Research Institute, Medical College of Wisconsin
[3]Great Lakes Water Institute, University of Wisconsin-Milwaukee
[4]Great Lakes Water Institute, University of Wisconsin-Milwaukee

Abstract

Detailed information about the prevalence of bacterial taxa in water samples can be determined via next generation sequencing. Samples were collected from two different, but related, water treatment sites in Lake Michigan over a three year period. After aggregation of similar taxa, there were 22 time series of bacterial prevalence data for each of the two sites.

The goal was to identify the taxa that had similar temporal patterns. Wavelet analysis after filtering to reduce sampling noise was used to determine the temporal characteristics of each of the time series. Wavelets were chosen as a functional analysis tool because of the irregular time measurements and the jagged shape of the prevalence curve. A discussion of the various problems that occur with collecting time series over long periods of time in an environmental study is presented along with some of the alternative solutions.

Comparisons of clustering methods of the wavelet coefficients and the similarities and differences of the corresponding temporal patterns show the stability and utility of this method for dealing with temporal and ordinal data.

Key Words: Genetics, Wavelet, Cluster Analysis, Fourier Decomposition

1. Introduction

The Milwaukee Metropolitan Sewerage District encompasses part or all of 6 watersheds and spans 1065 km^2 with 28 communities and goes to 2 Waste Water Treatment plants. It consists of two separate treatment plants: JI and SS. JI services a separated residential/industrial system as well as a combined sewer system in the oldest, most urbanized area of the city (green), SS WWTP primarily processes residential sewage (yellow). Water samples were collected over a 3 year period from each pipe and each sample was sequenced to identify the prevalence of the bacterial taxa. Using massively parallel pyrosequencing, we generated more than 1 million pyrotag sequences from the V6 hypervariable region of bacterial 16S rRNA genes from 19 paired wastewater influent samples from two plants and two samples taken upstream in the sanitary sewer system.

Comparison to a previously published human intestinal dataset revealed the majority of influent taxa to be of non-fecal origin. *Acinetobacter*, *Trichococcus*, and *Aeromonas* were at low abundance in the human and estuary samples, yet accounted for nearly 35% of the total sewage community.

Next generation DNA sequencing provides in depth description of microbial communities. **The complexity of these data sets challenges our ability to place these data into an ecological context. It also presents major statistical challenges which will be addressed in this paper.**

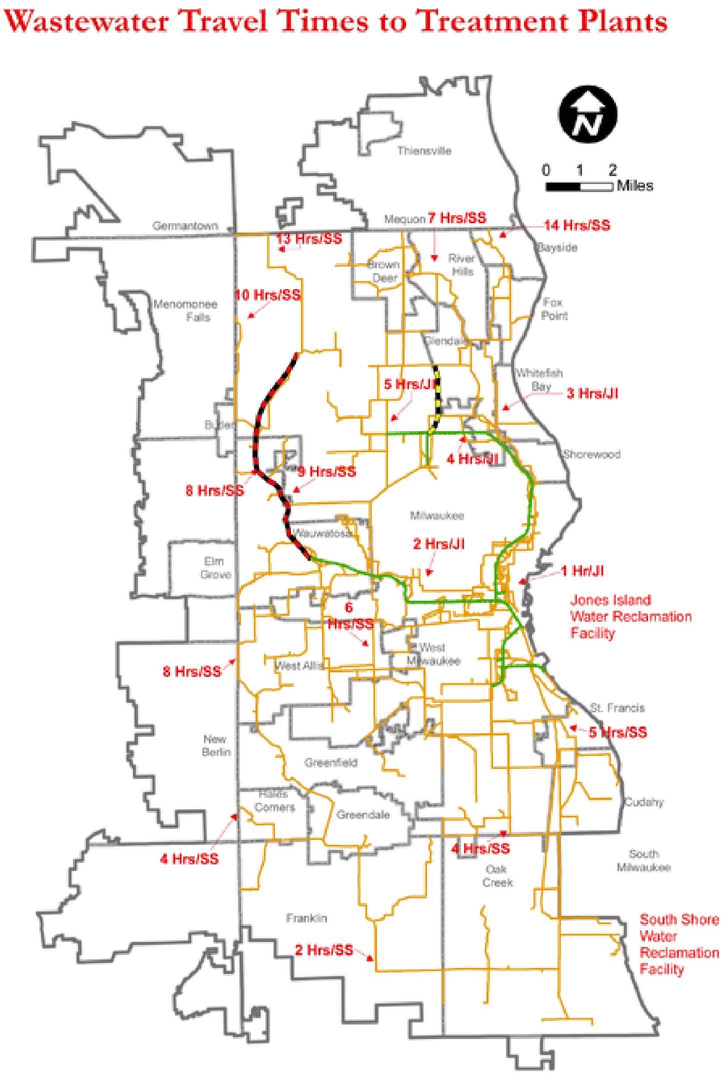


Figure 1: The two urbans sewer structures and their WWTPs, JI and SS.

1.1 Details About the Data

For example, while the Acinitobacter as a whole (75 different sub taxa) did not vary more than 20% in each sample, the individual taxa showed substantial seasonal effects that were not time concordant. Since the study was conducted in Wisconsin, no samples were collected in the winter months when the shore of Lake Michigan was frozen over.

Influent Sampling 19 paired sewage influent samples were collected from JI and SS WWTPs (n=38 total) over a three-year period. 102 taxa of the 1057 taxa were observed to account for 95% of the data. 18 non-taxa accounted for 66% of the pyrotags. 4 human taxa dominated. 3 taxa dominated the sewage samples, accounted for 33.6% of all pyrotags and are used to illustrate the process.
Samples consisted of 1 L of 24-hour flow-weighted samples collected from 6 am on the preceding day until 6 am on the stated collection day. Flow into the WWTP was averaged between measurements taken at the 6 am time points each day. Meteorological data accompanying the samples included high/low temperatures (five day average of collection day and four days previous) and precipitation totals (the day of and for 48 hrs previous). Ancillary data collected at the WWTPs included flow, ammonia determined by SM(20) 4500-NH3D, and BOD (5 day total)

The number of dominant Taxa consists of 18 sewage and 4 human giving a total of 22 concurrent time series of irregularly collected data. The bacterial taxa consist of the following:

Non-human

- Acinito tag 1
- Acinito tag 2
- Sum of the 73 remaining Acinito taxa
- Acidovorax Sum 26
- Aeromonas Sum 56
- Aeromonas Tag 1
- Aeromonas Tag 2
- Arcobacter Sum 22
- Bacteroidetes Sum 27
- Commamonadaceae Sum 44
- Enterobacteriaceae Sum 21
- Fusobacteriales Tag 1
- Lactococcus Sum 18
- Neisseriaceae Sum 6
- Psuedomonas Sum 41
- Simplicispira Sum 6
- Sporocytophaga Sum 16
- Trich Sum 21
- Trich Tag 1

Human:

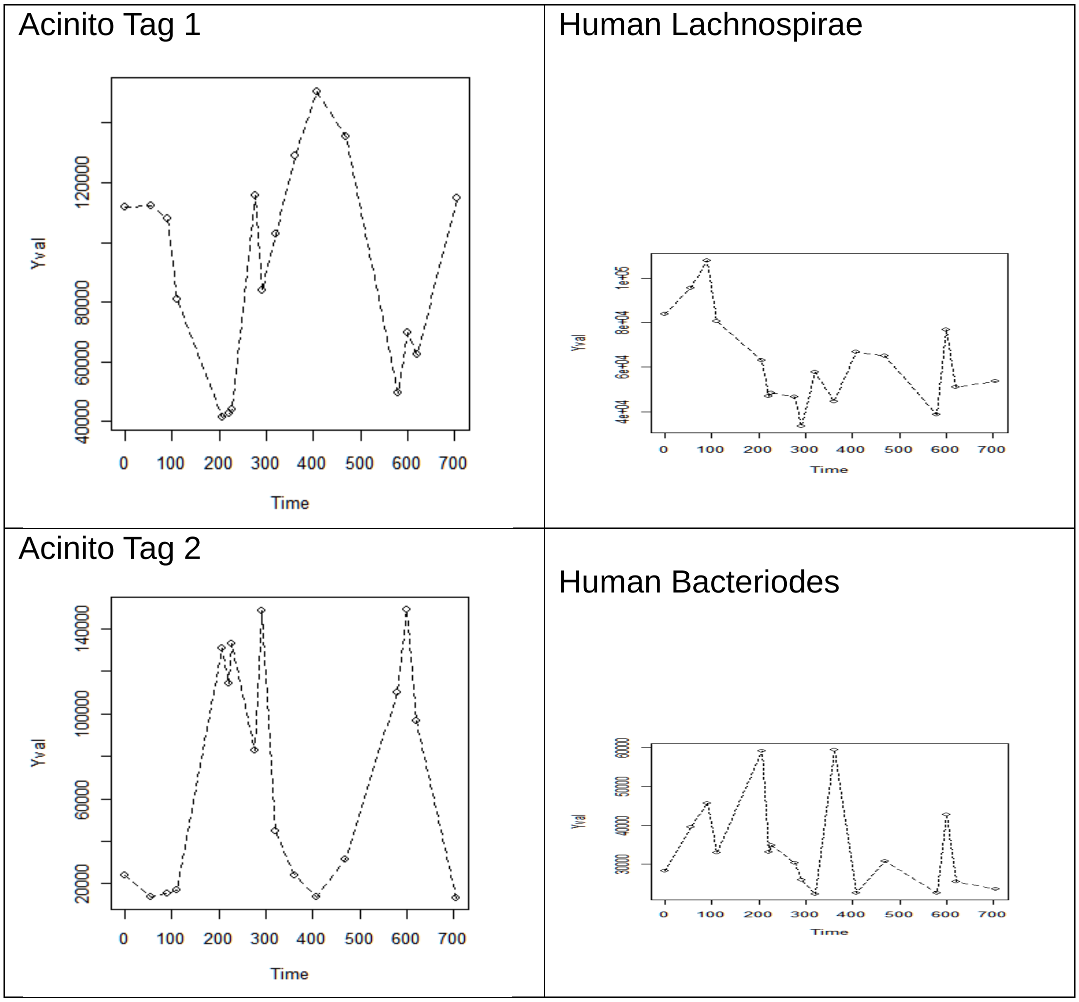
- Bacteroides
- Faecalibacterium
- Lachnospiraceae
- Parabacteroides

1.2 Goals of the Statistical Analysis

First, to examine the pattern of the bacterial prevalence over time, second to examine how the bacterial prevalence related to other environmental covariates: rainfall, high and low temperature, flow, ammonia, phosphorous, solids, BOD and third to compare the patterns of the bacterial prevalence between the pipes.

1.3 Examples of the Temporal Trajectory of the Taxa Prevalence

Two of the Acinitobacter non-human taxa and two of the human taxa are displayed in figure 2. Notice the difference in magnitude of the prevalence of the non-human and human taxa in figure 2. We can expect that because of the difference in magnitude that the human taxa may well have more random noise and consequently the correlations of the environmental co-variables with the human taxa will be less than the correlations among the time series correlations from the non-human taxa. Also note that the two Acinitobacter Tags are almost 180^o out of phase.



Acinito Tag 1

Human Lachnospirae

Acinito Tag 2

Human Bacteriodes

Figure 2: The Two Dominant Taxa over Time

2. Issues in Data Analysis:

This section discusses the statistical problems associated with trying to find similarities and differences in the temporal trend of the 18 non-human and the 4 human taxa. The first attempt at solving this problem was to use cluster analysis on the 19 point time series. Since clustering of the points does not consider the temporal order of the points, it was not successful. Clustering the 178 major taxa lead to one large cluster with about 100 members and 80 singleton clusters. Thus the goal was reduced to clustering the top 95% of the taxa – the 22 major time series. In order to preserve the temporal relationship of the data points, a functional model for the data was chosen that would describe the fluctuations in the prevalence over time.

One common solution to this problem for time series data is to use a harmonic series of sines and cosines at different frequencies. However, because the data was collected at irregular time points, it took 19 coefficients to model the irregular data and there was virtually no commonality among these coefficients across the major taxa.

A second solution to a functional representation of this data was to try to use a wavelet representation. The advantage of wavelets is that they are locally defined, making it easier to model irregular time series (Wavelet Methods in Statistics with R. G.P. Nason, Springer LLC. 2008) with short patterns and rapid changes.

2.1 Wavelet Analysis

The Daubechies wavelet (figure 3) is more complex than the Haar Wavelet, a step function, however it can both model sharp changes, and is mostly differentiable. It is also used by the R library “wavethresh” as the default wavelet for modeling.

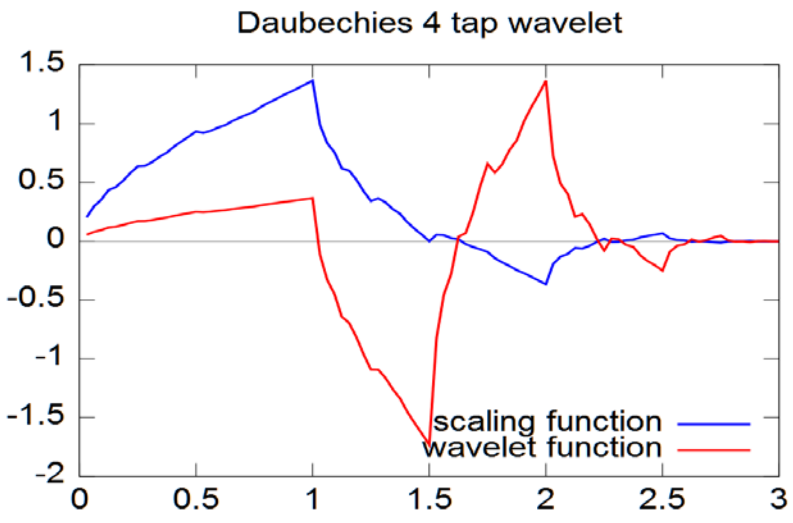


Figure 3: The Primary Daubechies Wavelet Function

This defines the "mother wavelet". The components of the wavelet transform are constructed by dilation and translation. Dilation makes the wavelet smaller, e.g. halving both its magnitude and the range over which it is non-zero. Translation moves these "baby" wavelets so that they cover the whole range of t's. For example the wavelet above would be subdivided into two babies; the first would go from 0 to 1.5 and the second from 1.5 to 3. Each would be scaled by a factor of 0.5. The succeeding levels of the wavelets would be halved in magnitude and range:

$$\boldsymbol{\theta_{j,k}(t) = 2^j\theta(2^j t - k)}$$

Since the $\theta_{j,k}$ form an orthonormal basis set, the decomposition of a function f(t) into

$$\boldsymbol{f(t)} = \sum_{\boldsymbol{0}}^{\boldsymbol{J}} \sum_{\boldsymbol{0}}^{\boldsymbol{K}} \propto_{\boldsymbol{j,k}} \boldsymbol{\theta_{j,k}(t)}$$

is analogous to a fast fourier transform. The $\propto_{j,k}$ are the wavelet coefficients of order j and location k. They are used to characterize the function f(t) and they allow clustering while preserving the temporal order. A final property of wavelets is that after fitting the function f(t), simple smoothing (or filtering) can be accomplished by reducing J to J-1, J-2, etc. The highest J corresponds to the finest detail of the wavelet representation**.**

An example of the filtering process can be seen in figure 4, where the first figure in the row is the raw data, the second figure is the smoothed data and the third figure is a plot of the wavelet coefficients, the $\propto_{j,k}$. The top most coefficient scales the mother wavelet over the interval, the second row gives the scaling of the two babies of order 2, etc.

The bottom row is the set of coefficients for J, the finest set of baby wavelets. The second row shows what the resulting curve looks like with less filtering/less smoothing of the data. While it is difficult to see, the lowest row has a larger magnitude in the second row - the coarsest filtering. The key information for comparing the wavelet functions; however, is in the upper rows, which contain the information on the general shape of the time series.

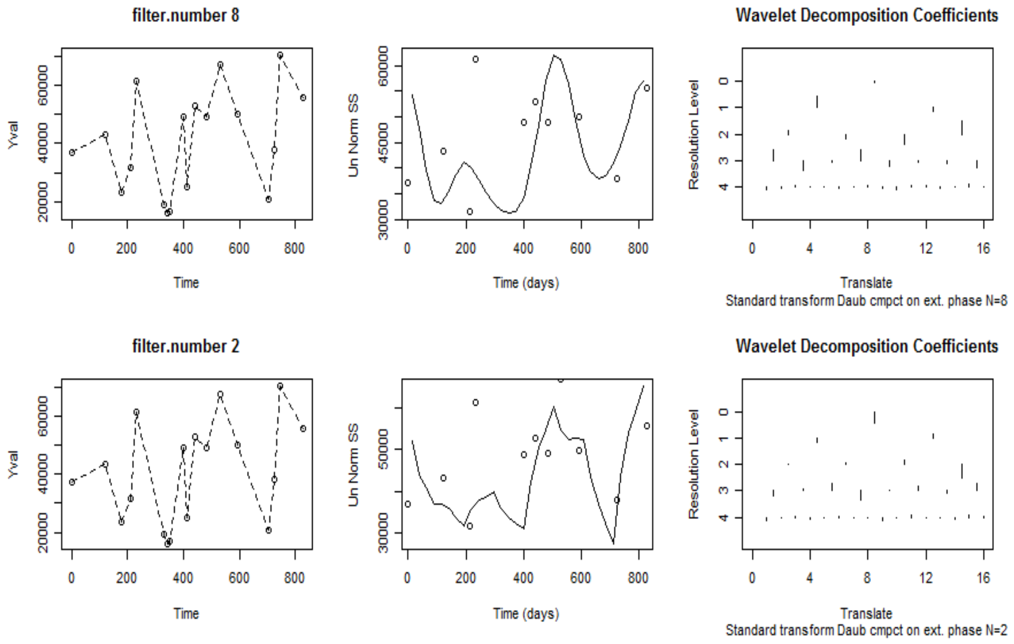


Figure 4: Wavelet Decomposition and Filtering

The analyses in this paper use the R library wavethresh. Wavethresh also can perform interpolation of the irregularly observed time series onto a regular series of points to aid the wavelet estimation process. We used this for all our wavelet decompositions.

2.2 Normalization of the samples

The first statistical analysis question is whether the different samples need to be normalized to make them comparable for analysis. The first method we used for normalization was to normalize the data for each pipe separately to one of the human bacteria (bacteriodes) which we thought would have relatively constant prevalence. The results of this normalization are shown in figure 5.

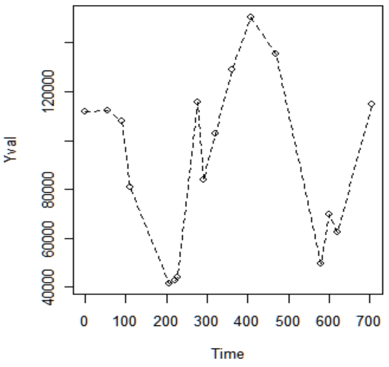
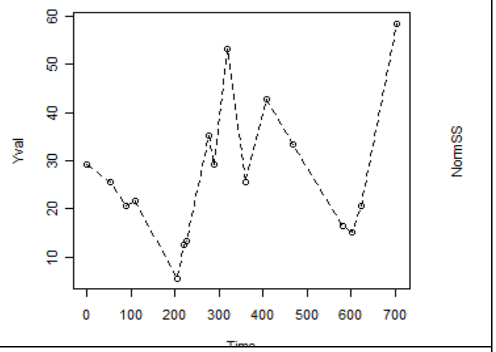
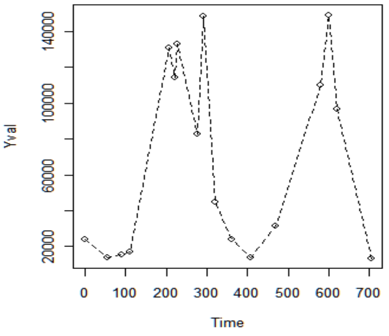
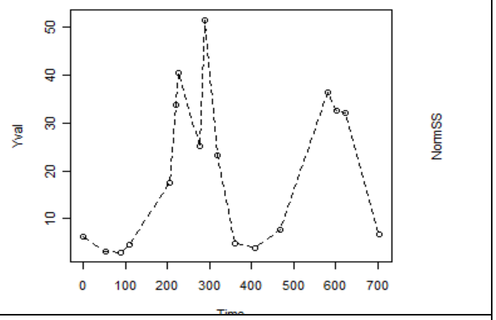
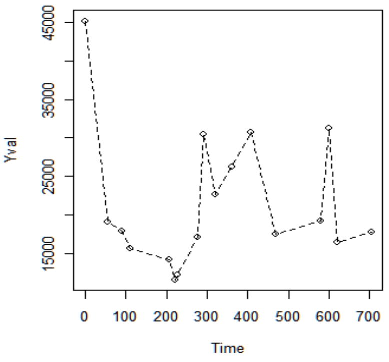
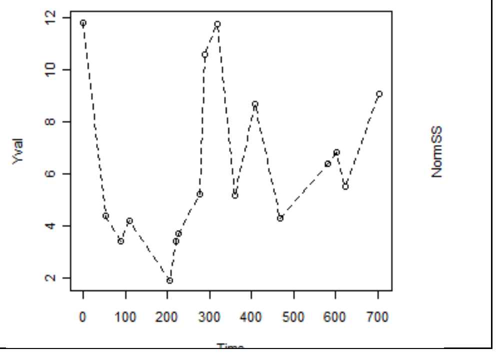
	Un-normalized versus normalized by the combined human bacteria	
Acinito Tag 1		
Acinito Tag 2		
Acinito Sum 73		

Figure 5: Normalization by Bacteriodes Prevalence

There are potentially other metrics for normalization: water flow in the pipe, overall prevalence, temperature. But most of these have a strong temporal fluctuation themselves and rather than normalizing, create residuals that correspond to what is not explained by that factor. Consequently, we used two approaches. Study the fluctuation in prevalence without normalization and study how much of the fluctuation is explained by the environmental variables.

2.3 Regression Analysis of the Environmental Variables

The environmental variables available consist of two general classes. The first class of variables is the global environmental variables:

- Rainfall for day -3
- Rainfall for day -2
- Rainfall for day -1 from the day when the sample was taken
- Daily high temp for each day from day -5 to day -1
- Daily low temp for each day from day -5 to day -1

The second class of environmental variables are specific to the pipe (water treatment plant) and consist of

- Flow in each pipe/WWTP
- Solids in each pipe/wwtp
- Ammonia level in each pipe/WWTP
- Phosphorus level in each pipe/WWTP
- BOD (biological oxygen delivery) level In each pipe

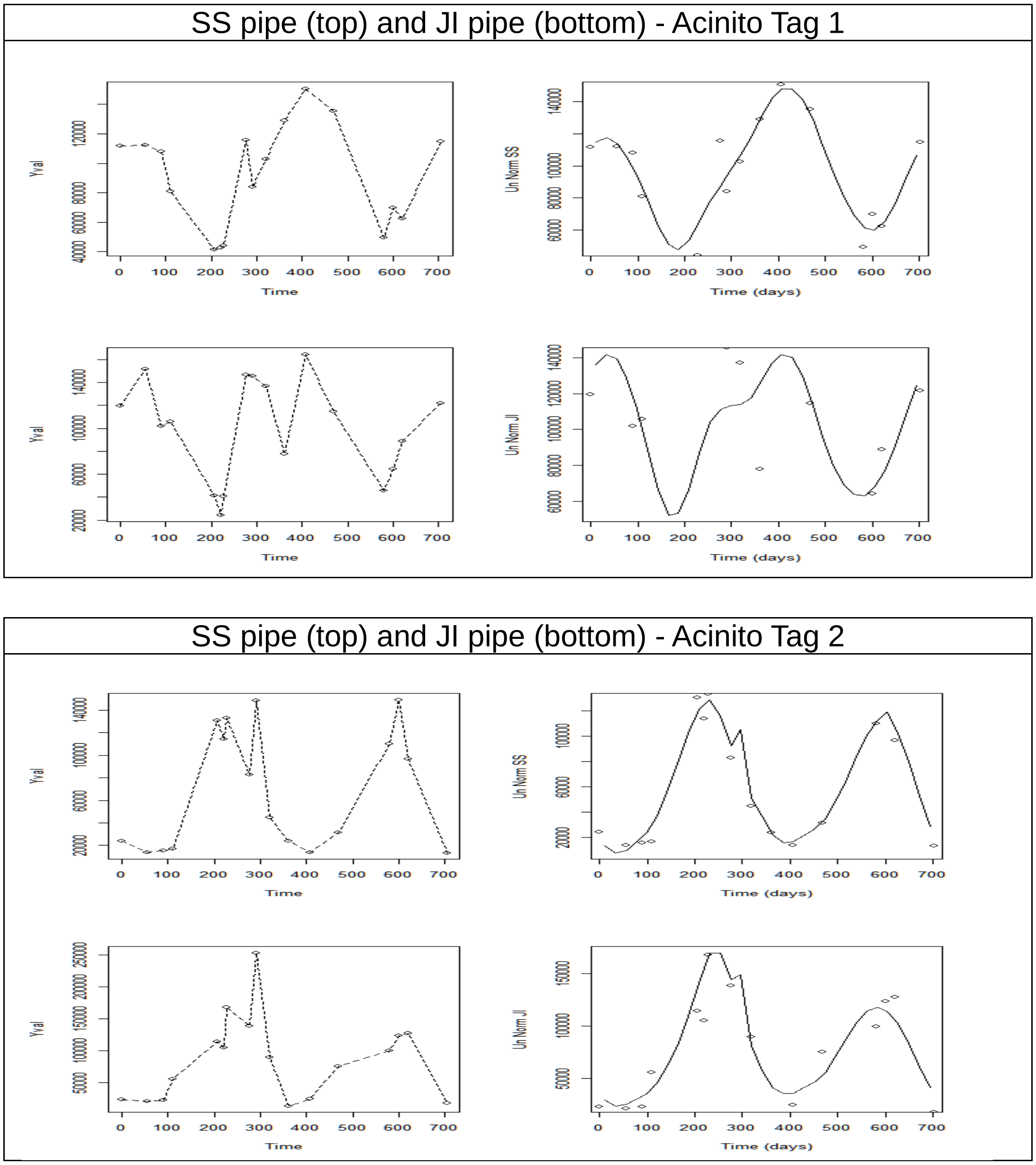
The environmental variables are not all independent. High and low temperatures have essentially the same temporal pattern, r = 0.98. Flow in each pipe is strongly related to phosphorous, ammonia, BOD and solids, r=0.76. Rainfall is independent of temperature and somewhat related to flow, r = 0.57.

In general, we observed that the Spearman Rank correlations of the environmental variables with the taxa were:

a. The correlations with the rainfall are generally low (not significant) for all the 18 taxa.
b. For the **human** taxa, the correlations between the measurements from **JI** and the temperatures (both low and high) are mostly significant; but the correlations for SS are not.
c. For the **human** taxa, only genus “**Parabacteroides”** has significant correlations with the other ancillary data (flow, ammonia, BOD, Phosphorus and Solids).
d. For the **non-human** taxa, the majority of them have highly significant correlations with the other ancillary data (flow, ammonia, BOD, Phosphorus and Solids).

2.3 Are the pipes different

The two pipes do represent different sources of water and may have somewhat different composition over time. As stated in the introduction, JI services a separated residential/industrial system as well as a combined sewer system in the oldest, most urbanized area of the city, SS WWTP primarily processes residential sewage. Figure 6 shows a comparison of the raw and smoothed data for the two dominant taxa. They are not identical, but very similar. Especially the wavelet filtered (smoothed) graphs on the right.



SS pipe (top) and JI pipe (bottom) - Acinito Tag 1

SS pipe (top) and JI pipe (bottom) - Acinito Tag 2

Figure 6: Comparison of the Two Pipes

As an additional measure of the similarity of the two pipes, we calculated the correlation between the pipes in table 1 for the non-human data and in table 2 for the human related bacteria.

Table 1 Non-Human Taxa

Tag Name	Correlation of JI and SS over time for non-normalized data
Acidovorax Sum 26 Tags	0.626 (p=0.006)
Acineto Sum 73 Tags	0.573 (p=0.013)
Acineto Tag 1	0.797 (p<0.001)
Acineto Tag 2	0.863 (p<0.001)
Aero Sum 56 Tags	0.770 (p<0.001)
Aero Tag 1	0.810 (p<0.001)
Aero Tag 2	0.794 (p<0.001)

Table 2 Human Related Taxa

Genus	Correlation for non-normalized data
Bacteroides	0.33762 (p=0.171)
Faecalibacterium	0.69021 (p=0.002)
Lachnospiraceae	0.57591 (p=0.012)
Parabacteroides	0.73766 (p<0.001)

The correlations for the non-human raw data are fairly high; somewhat less for the human data. This could be due to the different signal strength of the human taxa or due to differential sources for the water. the more noise and lower correlations. Rather than trying to devise and check some kind of weighting scheme, we considered each pipe separately and only report either the JI or the SS pipe results in the rest of this paper.

2.4 Interpreting the Pattern of Wavelet Coefficients:

Figure 7 shows the wavelet coefficients for two taxa that have quite different patterns over time. They are almost completely out of phase, which is easy to see from the highest level (lowest frequency) wavelet coefficients – large for Acinitobacter Tag 1 and small for Acinitobacter Tag 2. In addition the second and third level coefficient have much larger coefficients in the middle for the Acinitobacter Tag 2.

Figure 8 shows the corresponding wavelet coefficients for two somewhat similar taxa. While globally they agree as to when there uis higher prevalence – at the highest level of the wavelet coefficients - the differences in the shapes of the peaks is quite easily seen in the magnitude of the level 2 and level 3 coefficients. The advantage of the localization of the wavelet coefficients can easily be seen, compared to a sine-cosine harmonic regression model where any changes in shape affect all of the coefficients of the time serried decomposition. Even though in both cases, the bases are orthogonal.

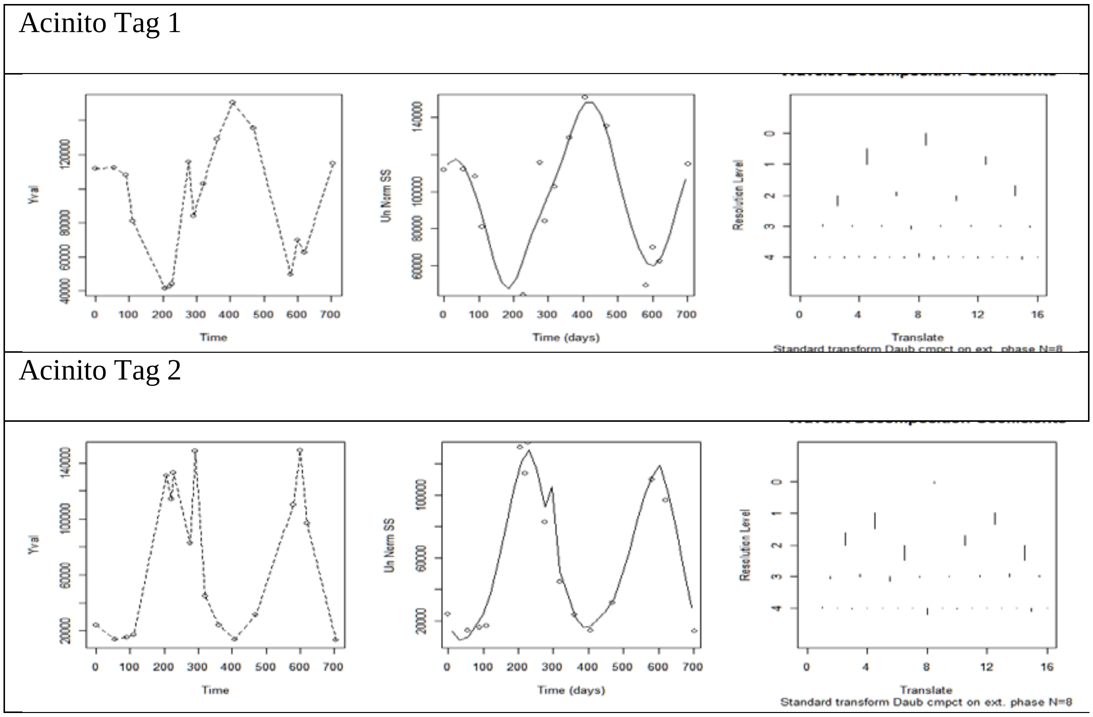


Figure 7: Different Patterns of Wavelet Coefficients corresponding to Time Series that are about 180^{o} out of phase

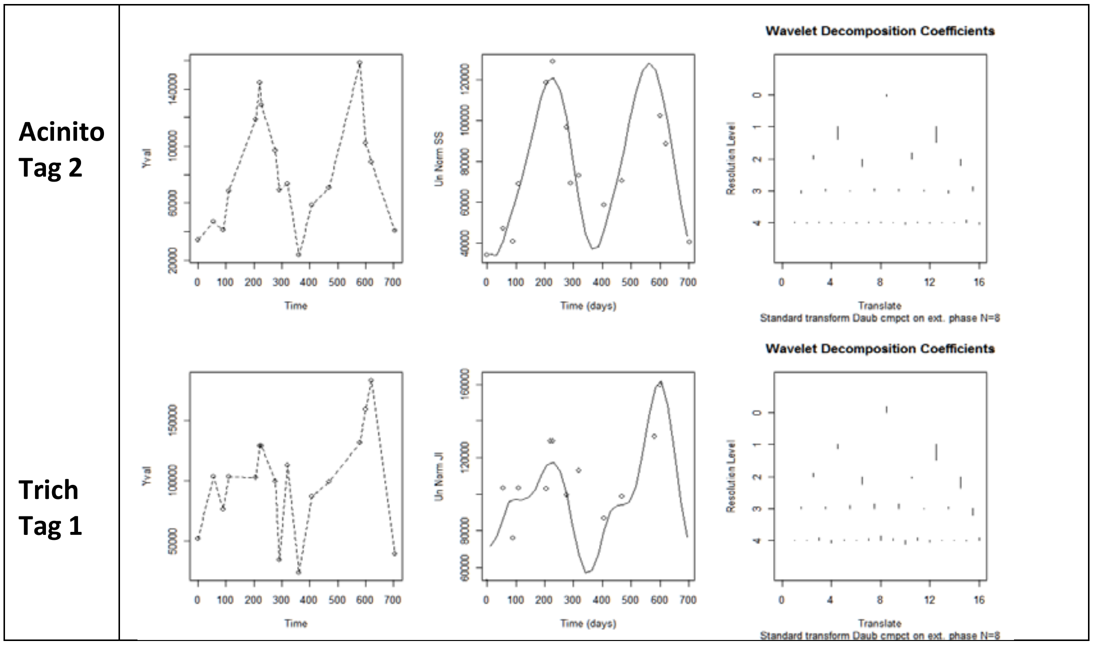


Figure 8: Similar Patterns of Wavelet Coefficients corresponding to similar Time Series

3. Clustering the taxa into similar temporal patterns

Hierarchical cluster analysis by itself, does not preserve the order of the data. However the wavelet coefficients should preserve the temporal structure.

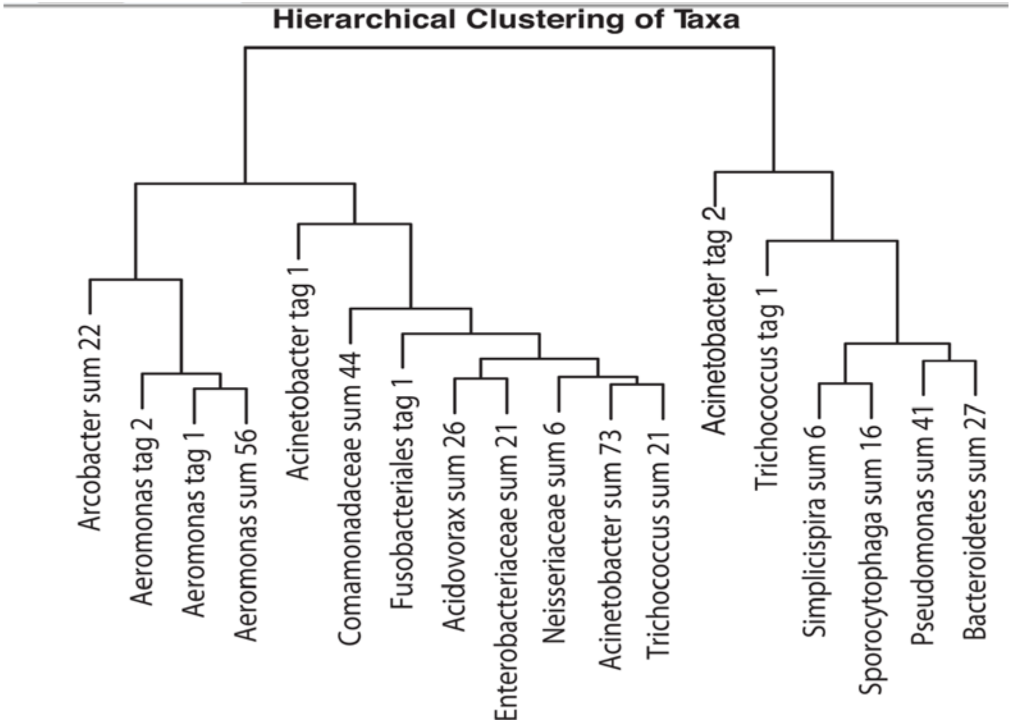


Figure 9: Clustering into Similar and Different Temporal Patterns

The cluster analysis of the wavelet coefficients in Figure 9 produces similar patterns to what we see visually when comparing the temporal patterns. For example, Acinobacter tag 1 and Acinobacter tag 2 are on opposite branches. Acinobacter tag 2 and Trichococcus tag 1 are nearby, although on different splits.

A concern is whether the choice of the clustering method produces different results. The effect of 3 differing clustering methods: complete, Ward and Average hierarchical clustering can be seen in figures 10, 11 and 12. As expected different clustering methods have some effect on the clusters. However, the figures appear to be more different than they actually are. For example in all three of them WD1 and WD11, are clustered together, they are just located on a different side of the diagram in the Ward clustering. WD5, WD8, WD12 and WD16 have the same set of branches in all three. Similarly for WD5, WD4 and WD7, etc. Thus as hoped and expected, the clustering of the temporal patters by representing them through wavelet coefficients is essentially robust to the method clustering.

Figure 10 Complete Clustering Method

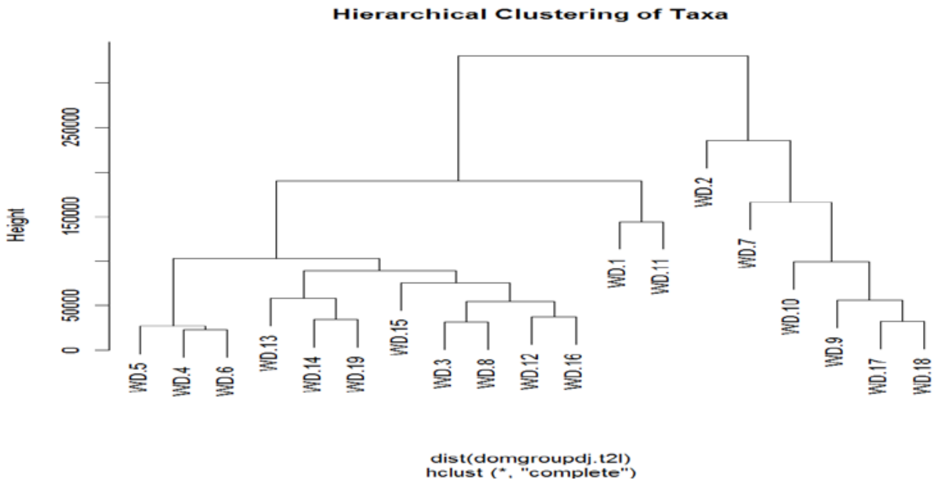


Figure 11 Ward Clustering Method

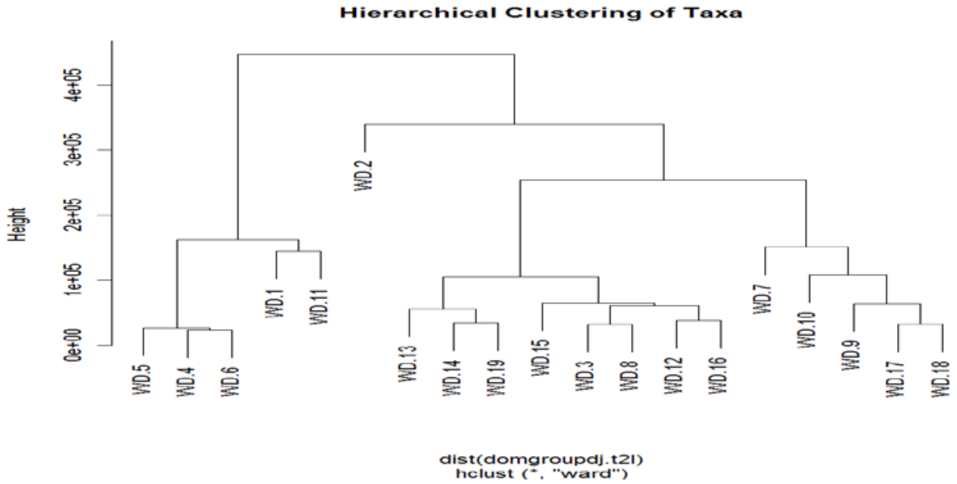
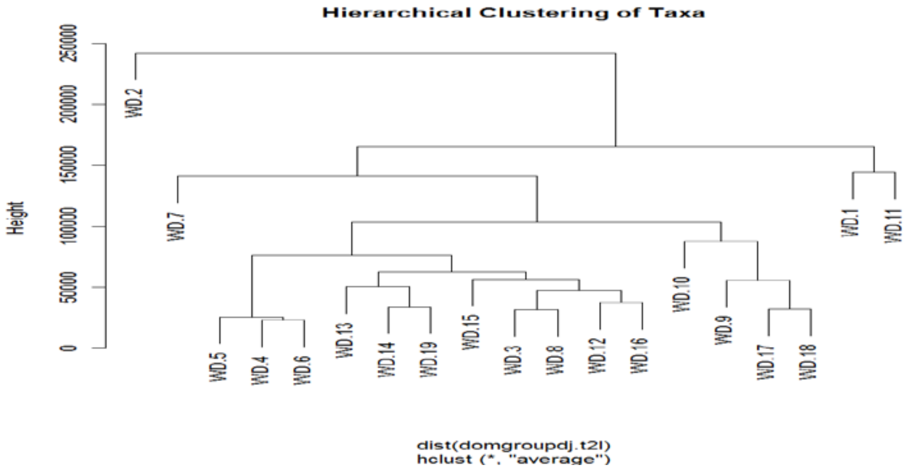


Figure 12 Average Clustering Method



References

1. Nason, G. P. Wavelet Methods in Statistics with R, Springer LLC. 2008 (Best wavelet reference as well as the library wavethresh and numerous examples)
2. Adler, J. R in a Nutshell. .O'Reilly, Inc. 2010 (Best R overview reference)
3. Statistics: An Introduction using R. Michael Crawley. John Wiley & Sons. 2005 (Both Stats and R)
4. The R Book. Michael Crawley. John Wiley & Sons. 2007 (An encyclopedic set of examples of using R)
5. On-line R references: http://cran.r-project.org/
6. R-commander: An elementary menuing system for R (For beginners to R this is a good starting point that also shows the R commands that are used)
7. Torrence C and Compo GP. "A Practical Guide to Wavelet Analysis",Bulletin of the American Meteorological Society, V98, 61-78, 1998.
8. Jenkins, G. M., and D. G. Watts, 1968: *Spectral Analysis and ItsApplications.* Holden-Day, 525 pp.
9. Chatfield, C., 1989: *The Analysis of Time Series: An Introduction.* .4th Ed. Chapman and Hall, 241 pp.
10. Daubechies, I., 1990: The wavelet transform time-frequency localization and signal analysis. *IEEE Trans. Inform. Theory,* **36,** 961–1004.
11. ——, 1992: *Ten Lectures on Wavelets.* Society for Industrial and Applied Mathematics, 357 pp.